Instruction-Following Evaluation for Large Language Models

Year: 2023 | Citations: 517 | Authors: Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu

Abstract

One core capability of Large Language Models (LLMs) is to follow natural language instructions. However, the evaluation of such abilities is not standardized: Human evaluations are expensive, slow, and not objectively reproducible, while LLM-based auto-evaluation is potentially biased or limited by the ability of the evaluator LLM. To overcome these issues, we introduce Instruction-Following Eval (IFEval) for large language models. IFEval is a straightforward and easy-to-reproduce evaluation benchmark. It focuses on a set of"verifiable instructions"such as"write in more than 400 words"and"mention the keyword of AI at least 3 times". We identified 25 types of those verifiable instructions and constructed around 500 prompts, with each prompt containing one or more verifiable instructions. We show evaluation results of two widely available LLMs on the market. Our code and data can be found at https://github.com/google-research/google-research/tree/master/instruction_following_eval